

# HUST

**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



# **XÂY DỰNG HỆ THỐNG THU THẬP, XỬ LÝ, LƯU TRỮ VÀ TRỰC QUAN HÓA DỮ LIỆU LỚN – THỬ NGHIỆM VỚI DỮ LIỆU CÁC CHUYẾN TAXI Ở THÀNH PHỐ NEW YORK**

**Sinh viên thực hiện:** Đào Anh Quân – 20215631

**GVHD:** TS. Đinh Thị Hà Ly

**ONE LOVE. ONE FUTURE.**

## Nội dung

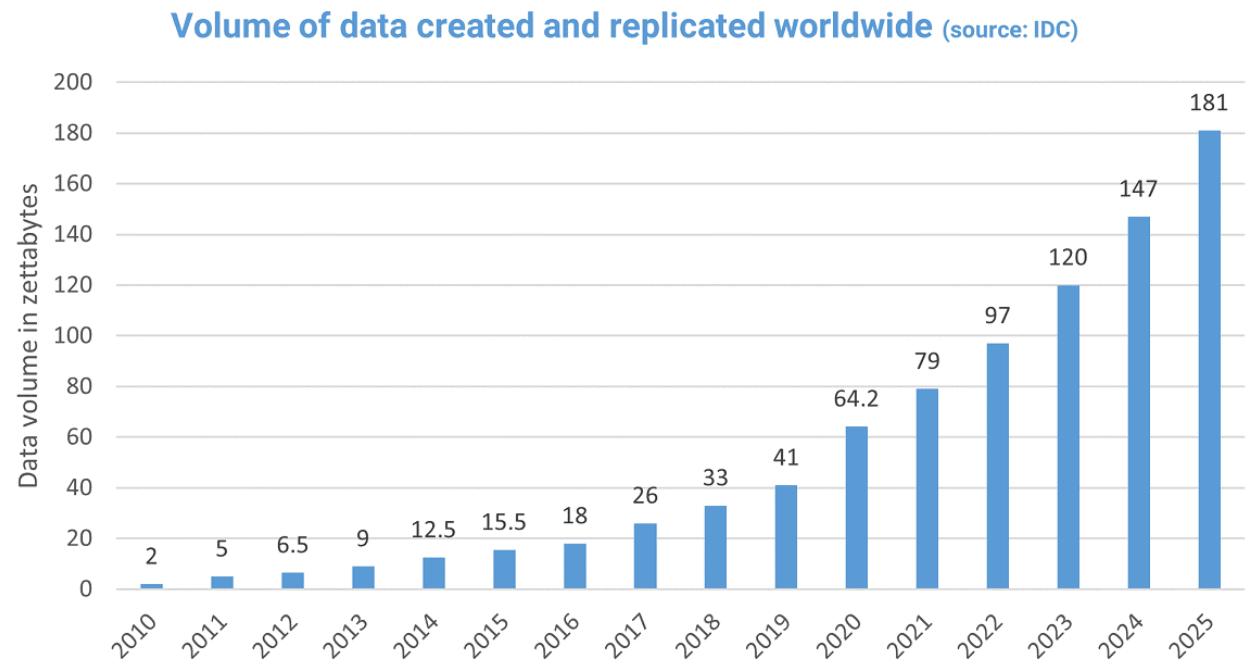
1. Giới thiệu đề tài
2. Thiết kế hệ thống
3. Triển khai hệ thống
4. Kết quả thực nghiệm
5. Kết luận và hướng phát triển

## Nội dung

- 1. Giới thiệu đề tài**
2. Thiết kế hệ thống
3. Triển khai hệ thống
4. Kết quả thực nghiệm
5. Kết luận và hướng phát triển

# 1. Giới thiệu đề tài

- **Đặt vấn đề:**
  - Dữ liệu ngày càng bùng nổ, khối lượng và tốc độ ngày càng tăng nhanh, mang nhiều giá trị tiềm năng.
  - Đặt ra yêu cầu xây dựng hệ thống phân tích dữ liệu lớn hoàn chỉnh, từ đó hỗ trợ ra quyết định nhanh chóng, chính xác và khách quan.



# 1. Giới thiệu đề tài

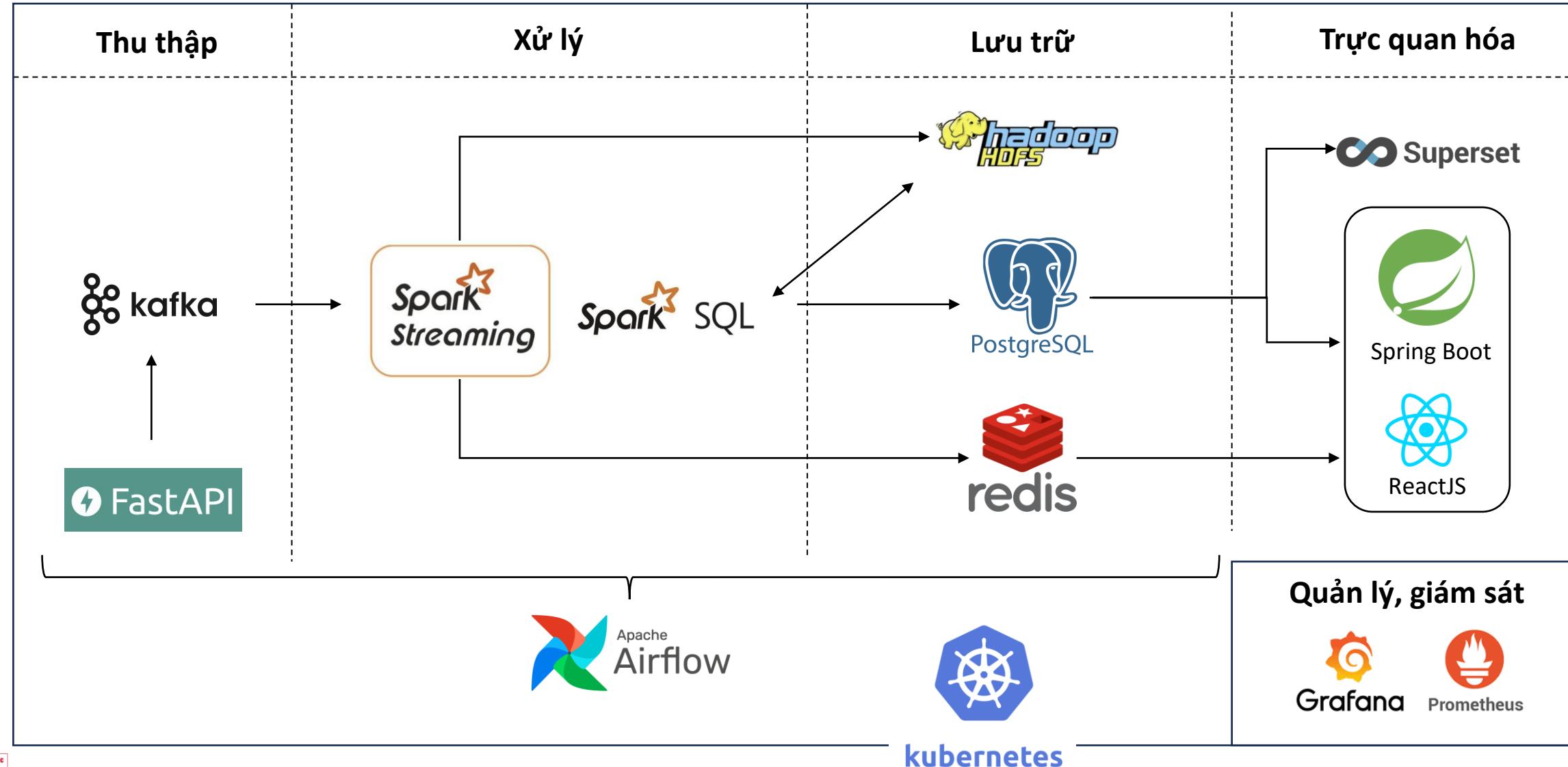
- Mục tiêu và giải pháp:

- Xây dựng một hệ thống xử lý dữ liệu lớn toàn chỉnh, bao gồm đầy đủ các mô-đun từ thu thập, xử lý, lưu trữ, trực quan hóa dữ liệu và quản lý giám sát.
- Kiến trúc Lambda, xử lý đồng thời dữ liệu theo luồng (streaming) và theo lô (batch). Triển khai trên môi trường Kubernetes.

## Nội dung

1. Giới thiệu đề tài
2. Thiết kế hệ thống
3. Triển khai hệ thống
4. Kết quả thực nghiệm
5. Kết luận và hướng phát triển

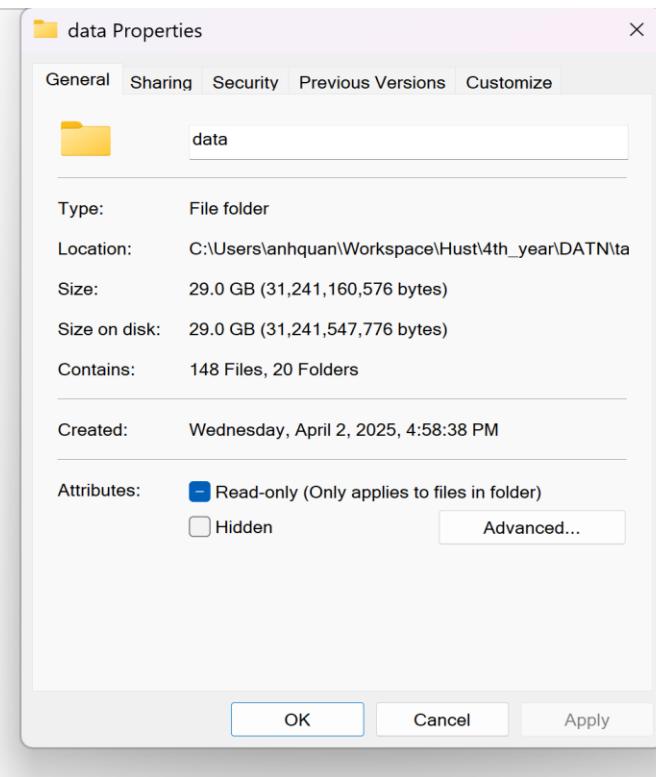
## 2. Thiết kế hệ thống



## 2. Thiết kế hệ thống

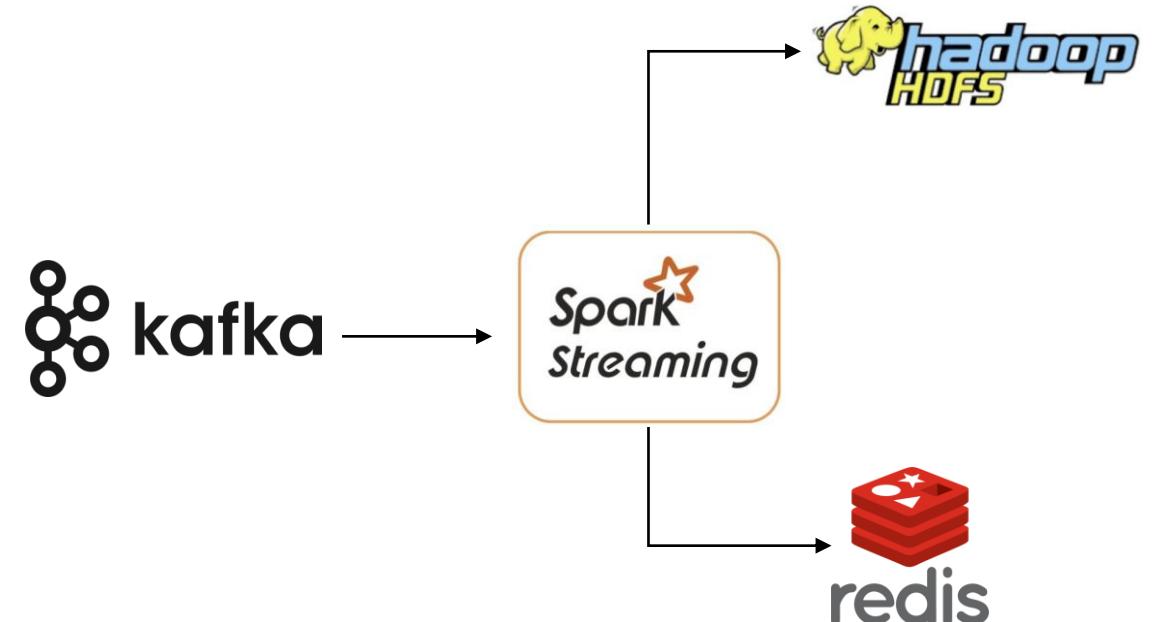
- Thu thập dữ liệu:
  - Giả lập thu thập dữ liệu qua API
  - Gửi đến hàng đợi thông điệp Kafka

Name	Date modified	Type	Size
yellow_tripdata_2019-01.parquet	6/1/2025 7:40 PM	PARQUET File	107,852 KB
yellow_tripdata_2019-02.parquet	6/1/2025 7:40 PM	PARQUET File	100,934 KB
yellow_tripdata_2019-03.parquet	6/1/2025 7:41 PM	PARQUET File	113,299 KB
yellow_tripdata_2019-04.parquet	6/1/2025 7:41 PM	PARQUET File	107,558 KB
yellow_tripdata_2019-05.parquet	6/1/2025 7:41 PM	PARQUET File	108,867 KB
yellow_tripdata_2019-06.parquet	6/1/2025 7:41 PM	PARQUET File	100,492 KB
yellow_tripdata_2019-07.parquet	6/1/2025 7:42 PM	PARQUET File	91,678 KB
yellow_tripdata_2019-08.parquet	6/1/2025 7:42 PM	PARQUET File	87,891 KB
yellow_tripdata_2019-09.parquet	6/1/2025 7:42 PM	PARQUET File	94,835 KB
yellow_tripdata_2019-10.parquet	6/1/2025 7:42 PM	PARQUET File	103,803 KB
yellow_tripdata_2019-11.parquet	6/1/2025 7:42 PM	PARQUET File	98,509 KB
yellow_tripdata_2019-12.parquet	6/1/2025 7:42 PM	PARQUET File	98,677 KB



## 2. Thiết kế hệ thống

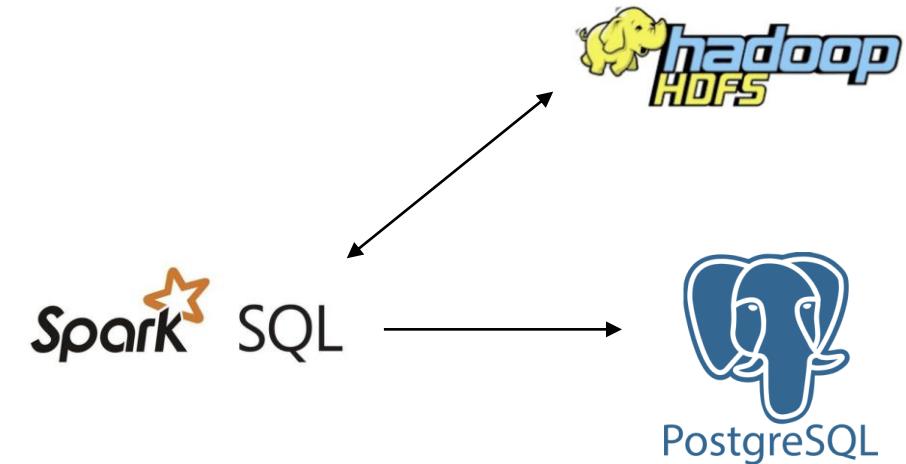
- Xử lý dữ liệu thời gian thực:
  - Đọc dữ liệu từ Kafka
  - Xử lý dữ liệu bằng Spark Streaming
  - Lưu dữ liệu thô trên HDFS
  - Dự đoán số chuyến đi theo thời gian thực, gửi đến Redis Pub/Sub



## 2. Thiết kế hệ thống

- Xử lý dữ liệu theo lô bằng Spark SQL:

- Thêm các trường dữ liệu về địa điểm
- Tính toán, phân tích các giá trị (thời gian, vận tốc trung bình,...)
- Loại bỏ dữ liệu bất thường
- Lưu dữ liệu trên HDFS, PostgreSQL
- Cập nhật mô hình dự đoán



## 2. Thiết kế hệ thống

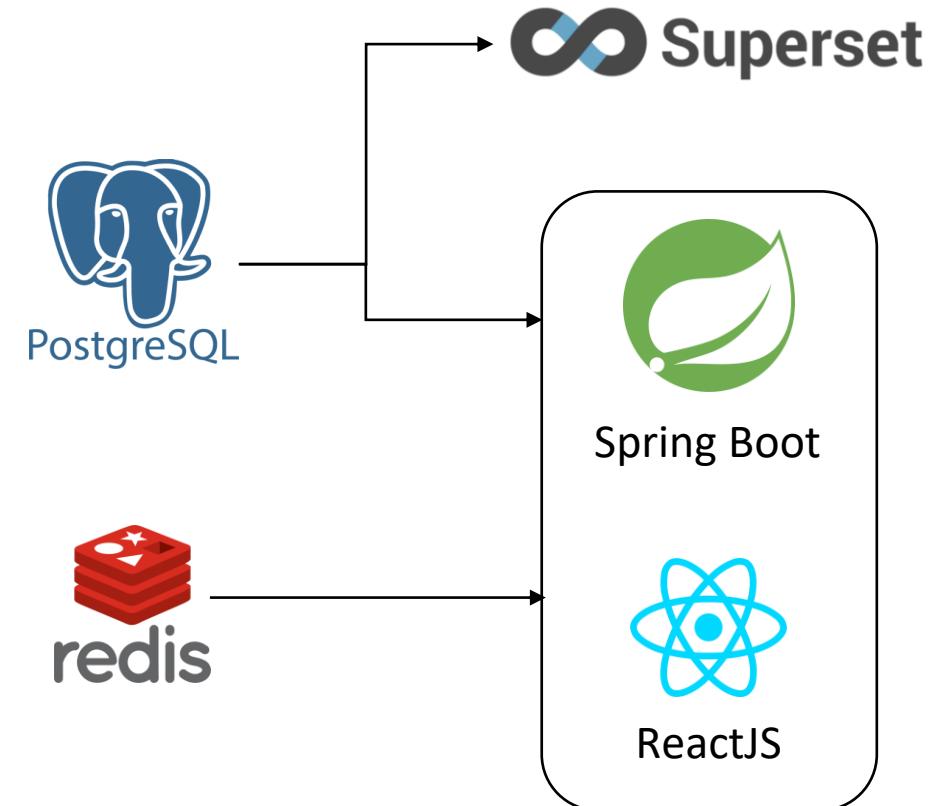
- Lưu trữ dữ liệu:
  - Dữ liệu checkpoints của quá trình xử lý streaming
  - Dữ liệu thô (dữ liệu lõi, hợp lệ)
  - Dữ liệu sau khi biến đổi, phân tích
  - Mô hình dự đoán
  - Các tài nguyên cần thiết khác



## 2. Thiết kế hệ thống

- Trực quan hóa dữ liệu:

- Tạo dashboard báo cáo phân tích tùy chỉnh
- Trực quan dữ liệu thời gian thực bằng Web Socket
- Bảng dữ liệu, biểu đồ phân tích cụ thể (theo thời gian, địa điểm)



## 2. Thiết kế hệ thống

- Quản lý, giám sát:

- Giao diện trực quan giám sát tài nguyên (RAM, CPU,...)
- Quản lý, điều phối hoạt động các thành phần



## Nội dung

1. Giới thiệu đề tài
2. Thiết kế hệ thống
3. Triển khai hệ thống
4. Kết quả thực nghiệm
5. Kết luận và hướng phát triển

### 3. Triển khai hệ thống

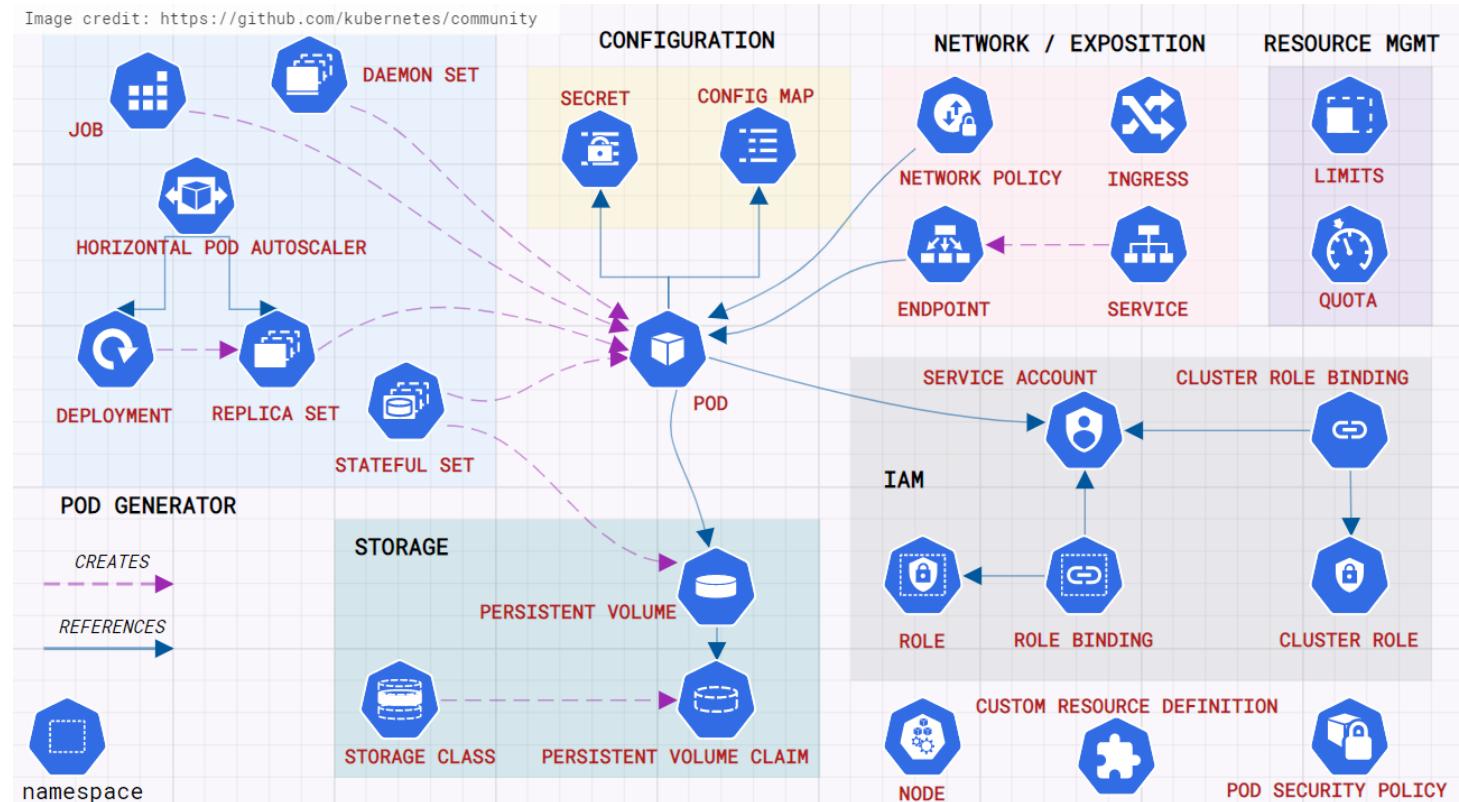
- Triển khai trên môi trường **K8s On-Premise**, bằng các máy ảo Linux:
  - 3 node chạy k8s đóng vai trò vừa là **control-plane**, vừa là **worker node**
  - 1 node cài đặt giao diện quản lý cụm k8s – **Rancher**
  - 1 node cài đặt **Ingress-Nginx**, thành phần cân bằng tải cho cụm
  - 1 node cài đặt nfs (Network File System), là nơi lưu trữ dữ liệu của các dịch vụ thông qua **PV (Persistent Volume)**

### 3. Triển khai hệ thống

- Cách thức triển khai:
  - Triển khai bằng Helm
  - Triển khai thủ công bằng các file yaml

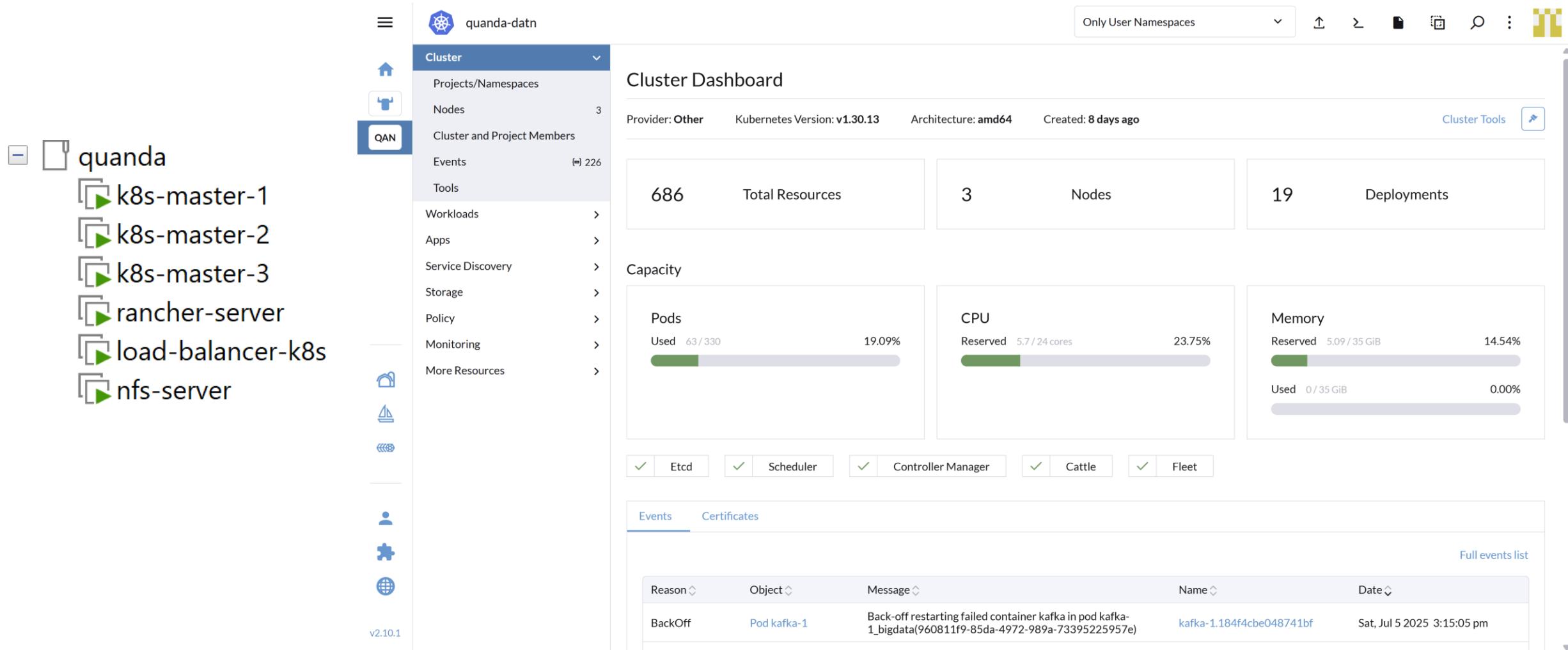


The  
package manager  
for Kubernetes



### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các node và giao diện Rancher

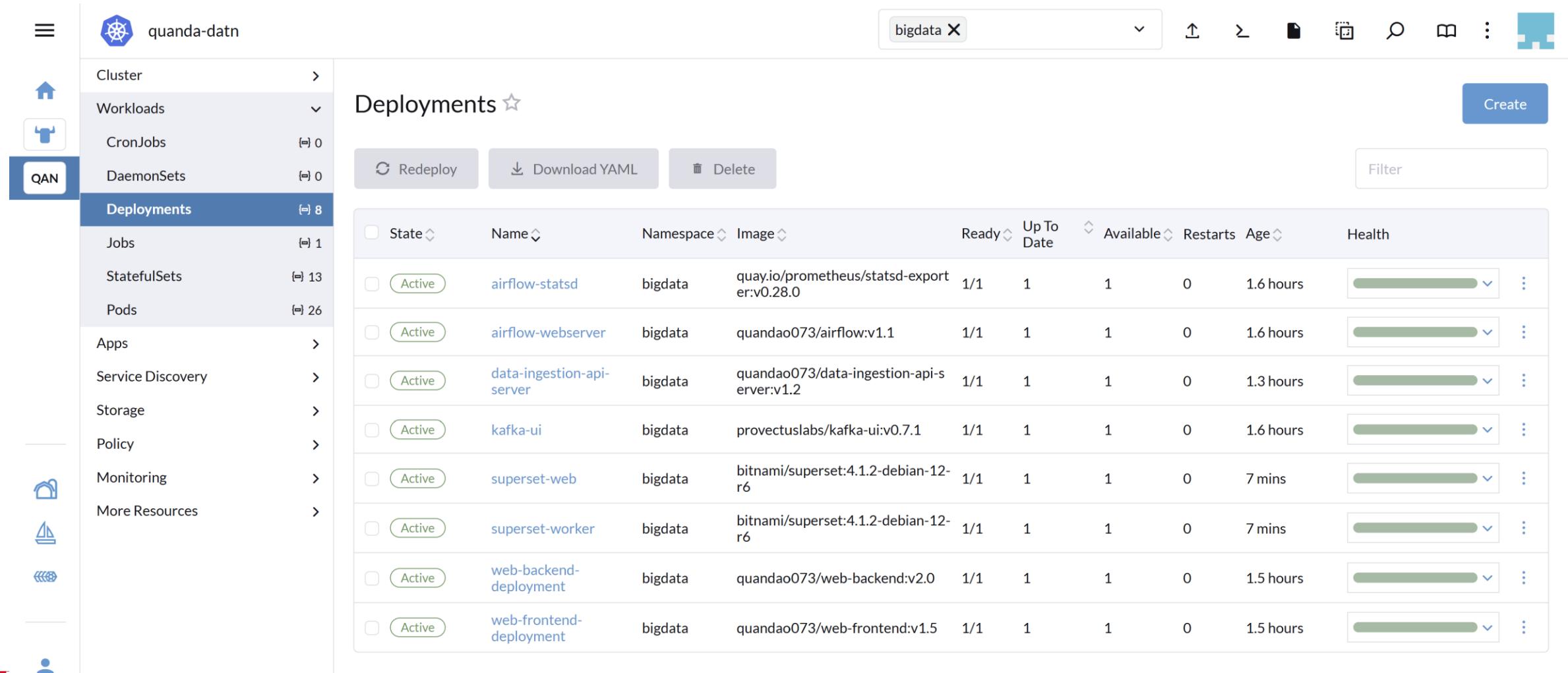


The screenshot shows the Rancher Cluster Dashboard for the 'quanda-data' cluster. The left sidebar lists the cluster's resources: 'quanda' (k8s-master-1, k8s-master-2, k8s-master-3), 'rancher-server', 'load-balancer-k8s', and 'nfs-server'. The main dashboard displays the following information:

- Cluster Dashboard:** Provider: Other, Kubernetes Version: v1.30.13, Architecture: amd64, Created: 8 days ago.
- Total Resources:** 686
- Nodes:** 3
- Deployments:** 19
- Capacity:**
  - Pods:** Used 63/330 (19.09%)
  - CPU:** Reserved 5.7/24 cores (23.75%)
  - Memory:** Reserved 5.09/35 GiB (14.54%)
  - Used 0/35 GiB (0.00%)
- Events:** Shows a recent event: 'BackOff' for 'Pod kafka-1' with the message 'Back-off restarting failed container kafka in pod kafka-1\_bigdata(960811f9-85da-4972-989a-73395225957e)'.

### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các Deployment

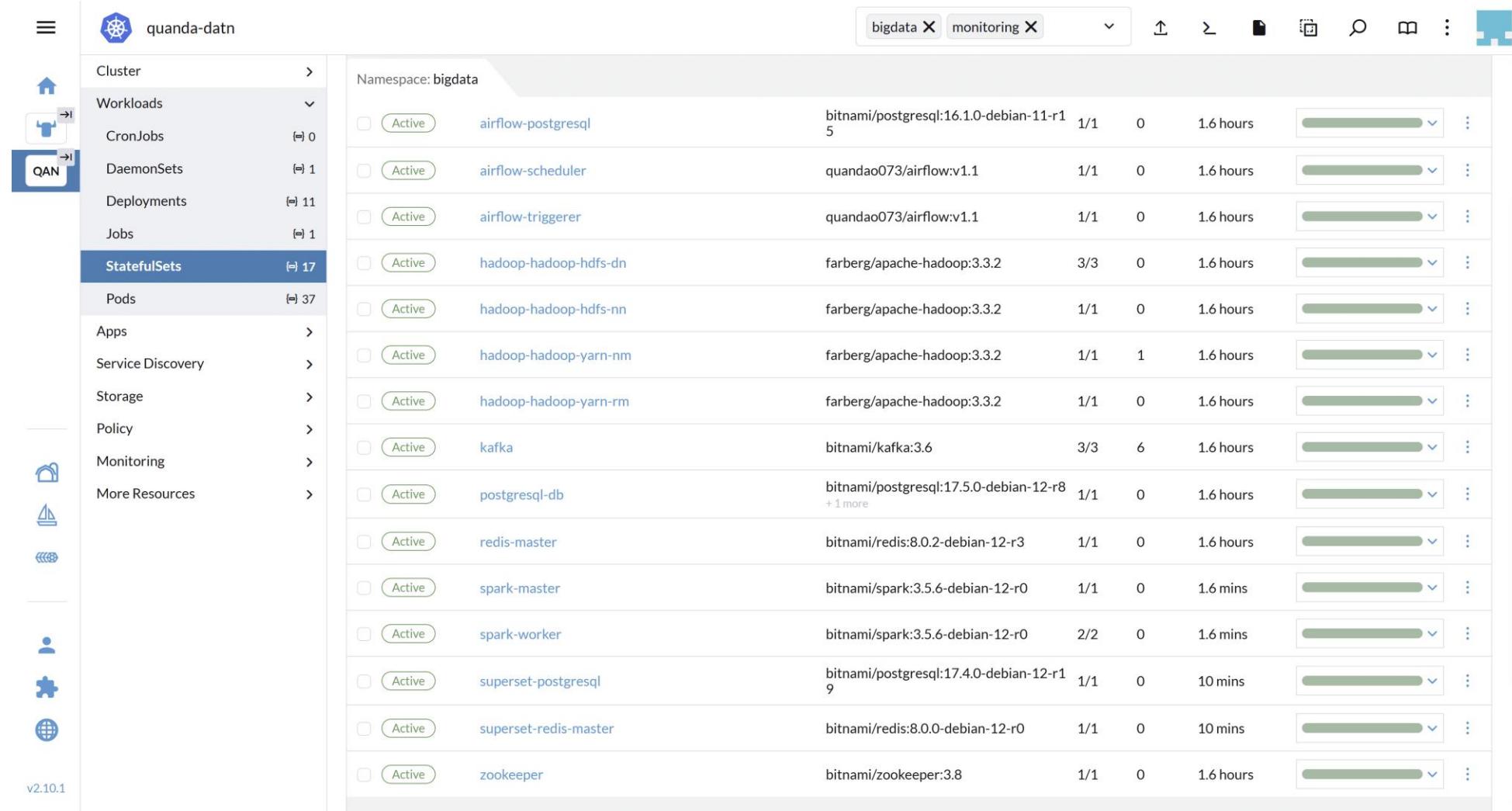


The screenshot shows the Quanda-DATN Kubernetes dashboard interface. The left sidebar has a 'QAN' tab selected, showing a list of resources: Cluster, Workloads (selected), CronJobs, DaemonSets, Deployments (selected, with 8 items), Jobs, StatefulSets (13 items), Pods (26 items), Apps, Service Discovery, Storage, Policy, Monitoring, and More Resources. The main area is titled 'Deployments' and shows a table with the following data:

State	Name	Namespace	Image	Ready	Up To Date	Available	Restarts	Age	Health
Active	airflow-statsd	bigdata	quay.io/prometheus/statsd-exporter:v0.28.0	1/1	1	1	0	1.6 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	airflow-webserver	bigdata	quandao073/airflow:v1.1	1/1	1	1	0	1.6 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	data-ingestion-api-server	bigdata	quandao073/data-ingestion-api-server:v1.2	1/1	1	1	0	1.3 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	kafka-ui	bigdata	provectuslabs/kafka-ui:v0.7.1	1/1	1	1	0	1.6 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	superset-web	bigdata	bitnami/superset:4.1.2-debian-12-r6	1/1	1	1	0	7 mins	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	superset-worker	bigdata	bitnami/superset:4.1.2-debian-12-r6	1/1	1	1	0	7 mins	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	web-backend-deployment	bigdata	quandao073/web-backend:v2.0	1/1	1	1	0	1.5 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>
Active	web-frontend-deployment	bigdata	quandao073/web-frontend:v1.5	1/1	1	1	0	1.5 hours	<div style="width: 100%;"><div style="width: 100%;"></div></div>

### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các StatefulSet



The screenshot shows the Quanda-DATN Kubernetes dashboard interface. The left sidebar has a 'StatefulSets' icon and is currently displaying the 'StatefulSets' section with 17 items. The main table lists 13 StatefulSets in the 'bigdata' namespace, all in an 'Active' state. The table columns include: Name, Image, Replicas, Ready, Available, Age, and a progress bar. The progress bars for most StatefulSets are nearly full, indicating they are fully deployed. The table also includes a 'More' column with three dots for each row.

Name	Image	Replicas	Ready	Available	Age	Progress Bar
airflow-postgresql	bitnami/postgresql:16.1.0-debian-11-r1	5	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
airflow-scheduler	quandao073/airflow:v1.1	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
airflow-triggerer	quandao073/airflow:v1.1	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
hadoop-hadoop-hdfs-dn	farberg/apache-hadoop:3.3.2	3	3/3	0	1.6 hours	<div style="width: 100%;"> </div>
hadoop-hadoop-hdfs-nn	farberg/apache-hadoop:3.3.2	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
hadoop-hadoop-yarn-nm	farberg/apache-hadoop:3.3.2	1	1/1	1	1.6 hours	<div style="width: 100%;"> </div>
hadoop-hadoop-yarn-rm	farberg/apache-hadoop:3.3.2	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
kafka	bitnami/kafka:3.6	3	3/3	6	1.6 hours	<div style="width: 100%;"> </div>
postgresql-db	bitnami/postgresql:17.5.0-debian-12-r8	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
redis-master	bitnami/redis:8.0.2-debian-12-r3	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>
						<div style="width: 100%;"> </div>
spark-master	bitnami/spark:3.5.6-debian-12-r0	1	1/1	0	1.6 mins	<div style="width: 100%;"> </div>
spark-worker	bitnami/spark:3.5.6-debian-12-r0	2	2/2	0	1.6 mins	<div style="width: 100%;"> </div>
superset-postgresql	bitnami/postgresql:17.4.0-debian-12-r1	1	1/1	0	10 mins	<div style="width: 100%;"> </div>
superset-redis-master	bitnami/redis:8.0.0-debian-12-r0	1	1/1	0	10 mins	<div style="width: 100%;"> </div>
zookeeper	bitnami/zookeeper:3.8	1	1/1	0	1.6 hours	<div style="width: 100%;"> </div>

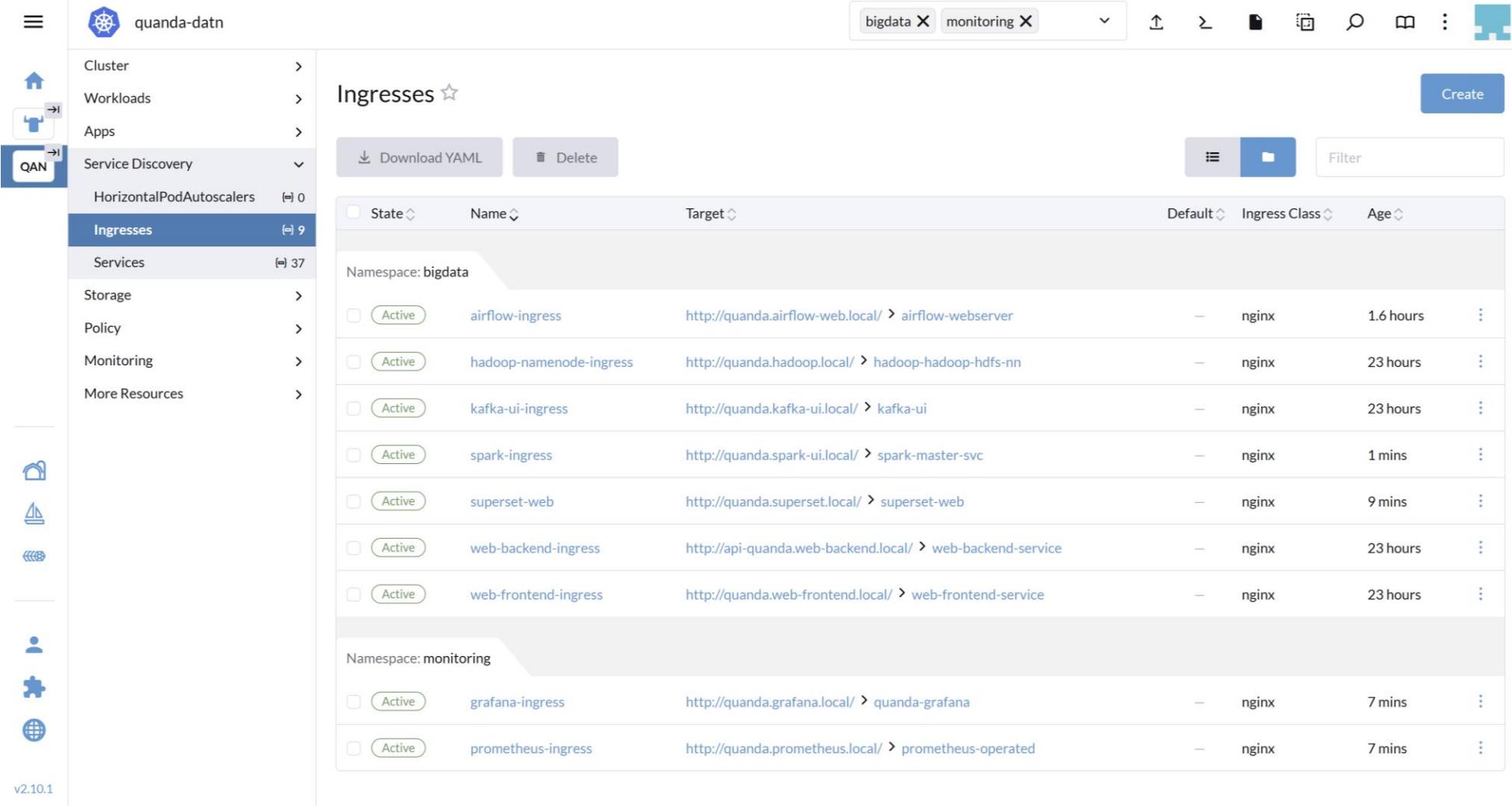
### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các Service

```
anhquan@k8s-master-1:~$ kubectl get service -n bigdata
NAME           TYPE      CLUSTER-IP   EXTERNAL-IP   PORT(S)          AGE
airflow-postgresql   ClusterIP  10.107.35.131 <none>        5432/TCP        58m
airflow-postgresql-hl ClusterIP  None          <none>        5432/TCP        58m
airflow-scheduler   ClusterIP  None          <none>        8793/TCP        58m
airflow-statsd      ClusterIP  10.107.145.66  <none>        9125/UDP,9102/TCP 58m
airflow-triggerer   ClusterIP  None          <none>        8794/TCP        58m
airflow-webserver   ClusterIP  10.105.9.245  <none>        8080/TCP        58m
data-ingestion-api-server ClusterIP  10.109.78.84 <none>        5000/TCP        22h
hadoop-hadoop-hdfs-dn ClusterIP  None          <none>        9000/TCP        50m
hadoop-hadoop-hdfs-nn ClusterIP  None          <none>        9000/TCP,9870/TCP 50m
hadoop-hadoop-yarn-nm ClusterIP  None          <none>        8088/TCP,8082/TCP,8042/TCP 50m
hadoop-hadoop-yarn-rm ClusterIP  None          <none>        8088/TCP        50m
hadoop-hadoop-yarn-ui ClusterIP  10.98.113.112 <none>        8088/TCP        50m
kafka             ClusterIP  None          <none>        9092/TCP        22h
kafka-ui          ClusterIP  10.97.216.4   <none>        80/TCP          22h
postgresql-db      ClusterIP  10.106.148.159 <none>        5432/TCP        37m
postgresql-db-hl   ClusterIP  None          <none>        5432/TCP        37m
postgresql-db-metrics ClusterIP  10.105.81.203 <none>        9187/TCP        37m
redis-headless     ClusterIP  None          <none>        6379/TCP        42m
redis-master       ClusterIP  10.98.239.153 <none>        6379/TCP        42m
spark-headless     ClusterIP  None          <none>        <none>          44m
spark-master-svc   ClusterIP  10.111.3.226  <none>        7077/TCP,80/TCP 44m
superset-postgresql ClusterIP  10.110.58.233 <none>        5432/TCP        26m
superset-postgresql-hl ClusterIP  None          <none>        5432/TCP        26m
superset-redis-headless ClusterIP  None          <none>        6379/TCP        26m
superset-redis-master ClusterIP  10.101.93.10  <none>        6379/TCP        26m
superset-web        ClusterIP  10.103.96.227 <none>        80/TCP          26m
web-backend-service ClusterIP  10.99.28.221  <none>        8080/TCP        16m
web-frontend-service ClusterIP  10.109.241.1  <none>        80/TCP          15m
zookeeper          ClusterIP  None          <none>        2181/TCP        22h
```

### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các Ingress



The screenshot shows the Quanda Data interface with the 'Ingresses' section selected in the sidebar. The main view displays two sections: 'Namespace: bigdata' and 'Namespace: monitoring', each containing a list of active Ingresses.

Namespace	Ingress Name	Target	Default	Ingress Class	Age
bigdata	airflow-ingress	http://quanda.airflow-web.local/ > airflow-webserver	—	nginx	1.6 hours
	hadoop-namenode-ingress	http://quanda.hadoop.local/ > hadoop-hadoop-hdfs-nn	—	nginx	23 hours
	kafka-ui-ingress	http://quanda.kafka-ui.local/ > kafka-ui	—	nginx	23 hours
	spark-ingress	http://quanda.spark-ui.local/ > spark-master-svc	—	nginx	1 mins
	superset-web	http://quanda.superset.local/ > superset-web	—	nginx	9 mins
	web-backend-ingress	http://api-quanda.web-backend.local/ > web-backend-service	—	nginx	23 hours
	web-frontend-ingress	http://quanda.web-frontend.local/ > web-frontend-service	—	nginx	23 hours
monitoring	grafana-ingress	http://quanda.grafana.local/ > quanda-grafana	—	nginx	7 mins
	prometheus-ingress	http://quanda.prometheus.local/ > prometheus-operated	—	nginx	7 mins

### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai các PersistentVolume và PersistentVolumeClaim

#### PersistentVolumes ☆

[Download YAML](#)

<input type="checkbox"/> State	Name	Reclaim Policy	Persistent Volume Claim
<input type="checkbox"/> Bound	airflow-code-pv	Retain	airflow-code-pvc
<input type="checkbox"/> Bound	airflow-dags-pv	Retain	airflow-dags-pvc
<input type="checkbox"/> Bound	airflow-datasource-pv	Retain	airflow-datasource-pvc
<input type="checkbox"/> Bound	airflow-logs-pv	Retain	airflow-logs-pvc
<input type="checkbox"/> Bound	data-source-pv	Retain	data-source-pvc
<input type="checkbox"/> Bound	monitoring-pv	Retain	prometheus-quanda-kube-prometheus-sta-prometheus-db-prometheus-quanda-kube-prometheus-sta-prometheus-0
<input type="checkbox"/> Bound	postgres-db-pv	Delete	postgres-db-pvc
<input type="checkbox"/> Bound	pvc-1a27efb8-ab0b-437a-a444-b931efb8c5e7	Delete	data-kafka-0
<input type="checkbox"/> Bound	pvc-1af1ca90-fa0c-48b4-a6d7-9765f1c99a31	Retain	dfs-hadoop-hadoop-hdfs-dn-0
<input type="checkbox"/> Bound	pvc-2bea5ab4-de30-4f58-ab63-53db898ff930	Delete	redis-data-superset-redis-master-0
<input type="checkbox"/> Bound	pvc-6bc3cda5-1a58-4993-be2f-3a0c230bc622	Retain	dfs-hadoop-hadoop-hdfs-dn-2
<input type="checkbox"/> Bound	pvc-67f559ea-ecc6-4f67-8085-a91853fed781	Retain	data-airflow-postgresql-0

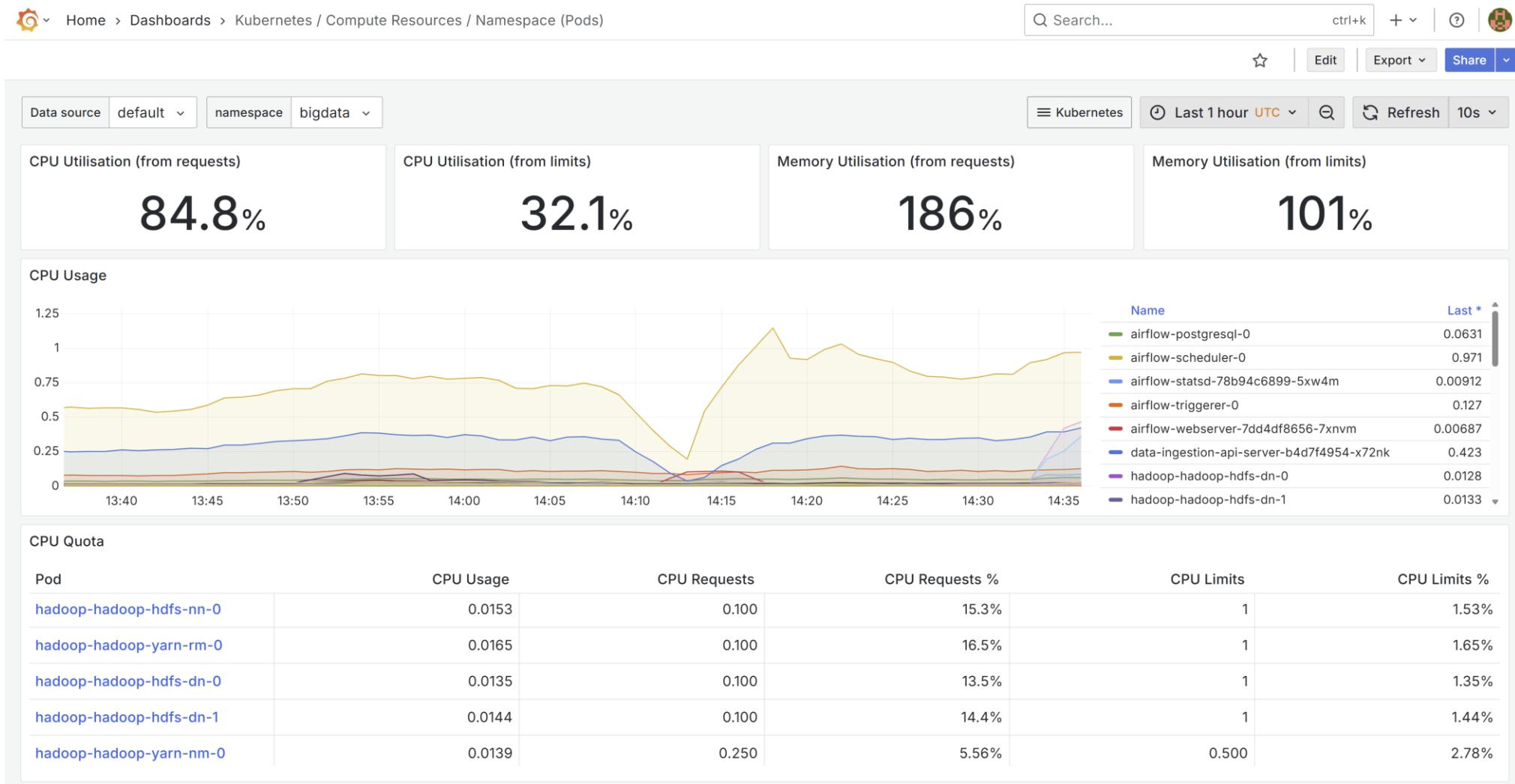
#### PersistentVolumeClaims ☆

[Download YAML](#) [Delete](#)

<input type="checkbox"/> State	Name	Namespace	Status	Volume	Capacity	Access Modes	Storage Class
<input type="checkbox"/> Bound	airflow-code-pvc	bigdata	Bound	airflow-code-pv	1Gi	RWX	nfs-storage
<input type="checkbox"/> Bound	airflow-dags-pvc	bigdata	Bound	airflow-dags-pv	1Gi	RWX	nfs-storage
<input type="checkbox"/> Bound	airflow-datasource-pvc	bigdata	Bound	airflow-datasource-pv	30Gi	RWX	nfs-storage
<input type="checkbox"/> Bound	airflow-logs-pvc	bigdata	Bound	airflow-logs-pv	5Gi	RWX	nfs-storage
<input type="checkbox"/> Bound	data-airflow-postgresql-0	bigdata	Bound	pvc-67f559ea-ecc6-4f67-8085-a91853fed781	8Gi	RWO	nfs-airflow-storage
<input type="checkbox"/> Bound	data-kafka-0	bigdata	Bound	pvc-1a27efb8-ab0b-437a-a444-b931efb8c5e7	5Gi	RWO	nfs-kafka-storage
<input type="checkbox"/> Bound	data-kafka-1	bigdata	Bound	pvc-493c8a08-2174-468d-8125-b3586affc54a	5Gi	RWO	nfs-kafka-storage
<input type="checkbox"/> Bound	data-kafka-2	bigdata	Bound	pvc-bd3e9605-f18f-4ac4-8daf-d2bd38e5df6e	5Gi	RWO	nfs-kafka-storage
<input type="checkbox"/> Bound	data-source-pvc	bigdata	Bound	data-source-pv	30Gi	ROX	nfs-storage
<input type="checkbox"/> Bound	data-superset-postgresql-0	bigdata	Bound	pvc-e75ecc80-5445-4c7b-9792-ff61dfac49ed	8Gi	RWO	nfs-superset-storage

### 3. Triển khai hệ thống (Kết quả triển khai)

- Triển khai công cụ giám sát tài nguyên



## Nội dung

1. Giới thiệu đề tài
2. Thiết kế hệ thống
3. Triển khai hệ thống
- 4. Kết quả thực nghiệm**
5. Kết luận và hướng phát triển

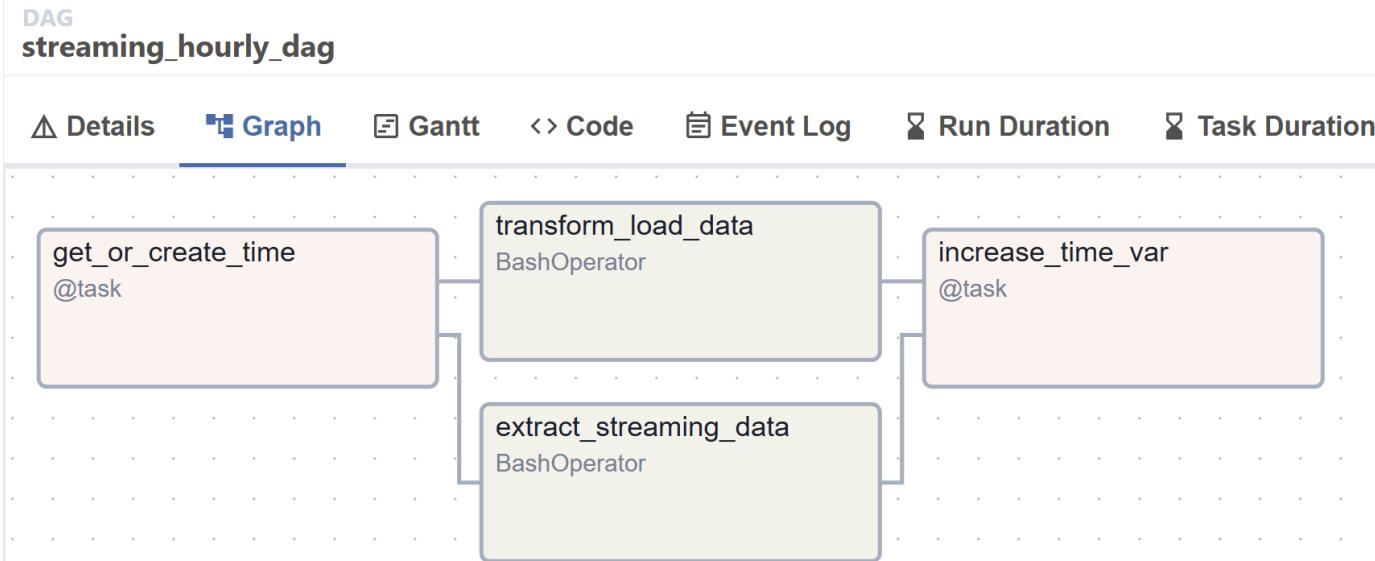
# 4. Kết quả thực nghiệm (Thu thập dữ liệu)

```
{  
  "status": "ok",  
  "data": [  
    {  
      "VendorID": 2,  
      "tpep_pickup_datetime": "2024-02-01T00:00:00",  
      "tpep_dropoff_datetime": "2024-02-01T22:55:05",  
      "passenger_count": 1,  
      "trip_distance": 5.25,  
      "RatecodeID": 1,  
      "store_and_fwd_flag": "N",  
      "PUlocationID": 68,  
      "DOlocationID": 264,  
      "payment_type": 2,  
      "fare_amount": 28.9,  
      "extra": 3.5,  
      "mta_tax": 0.5,  
      "tip_amount": 0,  
      "tolls_amount": 0,  
      "improvement_surcharge": 1,  
      "total_amount": 33.9,  
      "congestion_surcharge": 0,  
      "airport_fee": 0  
    },  
    {  
      "VendorID": 2,  
      "tpep_pickup_datetime": "2024-02-01T00:00:02",  
      "tpep_dropoff_datetime": "2024-02-01T00:25:07",  
      "passenger_count": 1,  
    }  
  ]  
}
```

Topics / yellow\_trip\_data

Overview	Messages	Consumers	Settings	Statistics	Type	Segment Size	Segment Count
Partitions 3	Replication Factor 3	URP 0	In Sync Replicas 9 of 9	Type External	Segment Size 12 MB	Segment Count 9	
Message Count 7393							
Partition ID	Replicas	First Offset	Next Offset	Message Count			
0	1, 2, 0	1589	3847	2258			
1	0, 1, 2	1422	3516	2094			
2	2, 0, 1	2049	5090	3041			

# 4. Kết quả thực nghiệm (Xử lý dữ liệu)



```
1) "message"
2) "realtime-trip-channel"
3) "{  
    \"timestamp\": \"2025-01-01 00:01:00\",  
    \"trip_count\": 11289,  
    \"predicted_timestamp\": \"2025-01-01 01:01:00\",  
    \"predicted_trip_count\": 16924  
}  
1) "message"
2) "realtime-trip-channel"
3) "{  
    \"timestamp\": \"2025-01-01 00:01:04\",  
    \"trip_count\": 11291,  
    \"predicted_timestamp\": \"2025-01-01 01:01:04\",  
    \"predicted_trip_count\": 16926  
}
```

# 4. Kết quả thực nghiệm (Xử lý dữ liệu)

DAG  
batch\_processing\_dag

△ Details     Graph     Gantt     Code     Event Log     Run Duration     Task Duration

get\_or\_create\_time  
@task

batch\_processing  
SparkSubmitOperator

update\_models  
SparkSubmitOperator

increase\_time\_var  
@task

DAG  
batch\_processing\_dag

△ Details     Graph     Gantt     Code     Event Log     Run Duration     Task Duration     Calendar

1

% Success

2023

0

Failed

2022

Success

2021

Failed

# 4. Kết quả thực nghiệm (Xử lý dữ liệu)



## Spark Master at spark://spark-master-0.spark-headless.bigdata.svc.cluster.local:7077

**URL:** spark://spark-master-0.spark-headless.bigdata.svc.cluster.local:7077

**Alive Workers:** 2

**Cores in use:** 6 Total, 0 Used

**Memory in use:** 10.0 GiB Total, 0.0 B Used

**Resources in use:**

**Applications:** 0 Running, 2 Completed

**Drivers:** 0 Running, 0 Completed

**Status:** ALIVE

### ▼ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20250602061720-172.16.196.49-39243	172.16.196.49:39243	ALIVE	3 (0 Used)	5.0 GiB (0.0 B Used)	
worker-20250602061803-172.16.168.8-41179	172.16.168.8:41179	ALIVE	3 (0 Used)	5.0 GiB (0.0 B Used)	

### ▼ Running Applications (0)

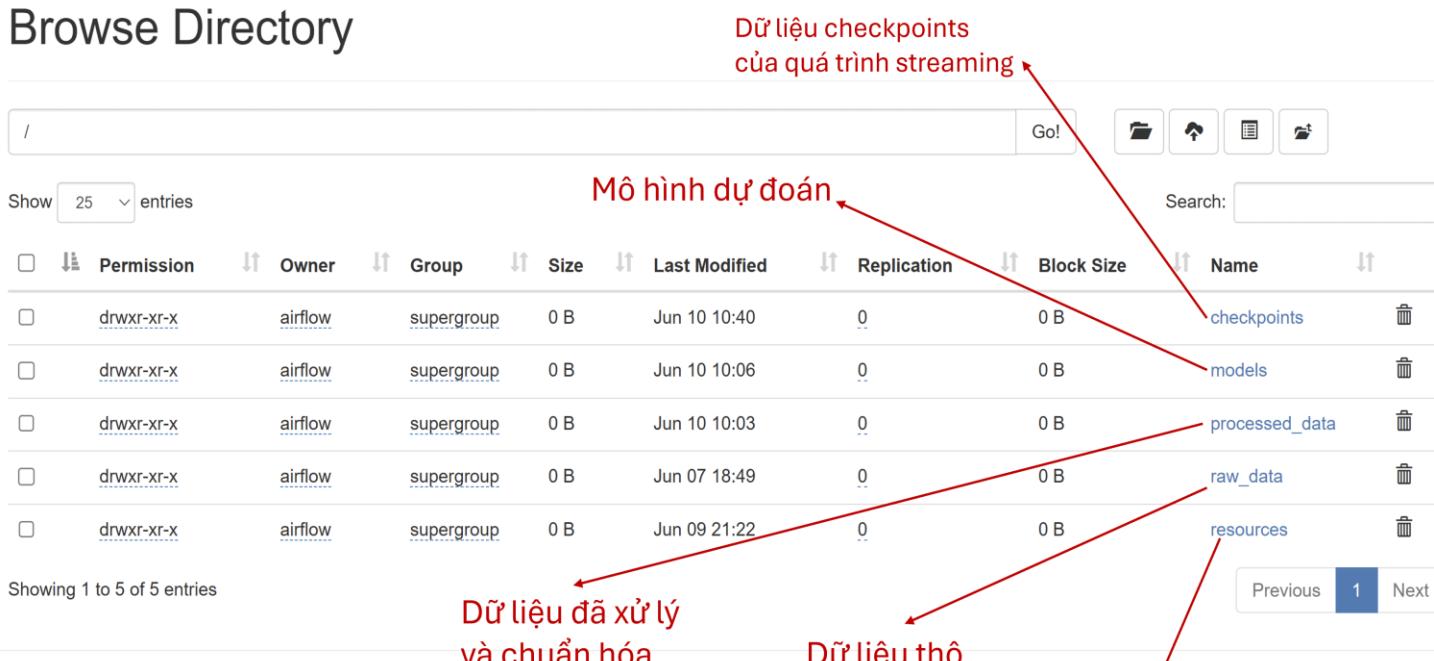
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

### ▼ Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20250602062921-0001	UpdateModel	6	1024.0 MiB		2025/06/02 06:29:21	airflow	FINISHED	16 s
app-20250602062351-0000	MonthlyBatchProcessing	6	1024.0 MiB		2025/06/02 06:23:51	airflow	FINISHED	5.4 min

# 4. Kết quả thực nghiệm (Lưu trữ dữ liệu)

## Browse Directory



Mô hình dự đoán

Dữ liệu checkpoints  
của quá trình streaming

Dữ liệu đã xử lý  
và chuẩn hóa

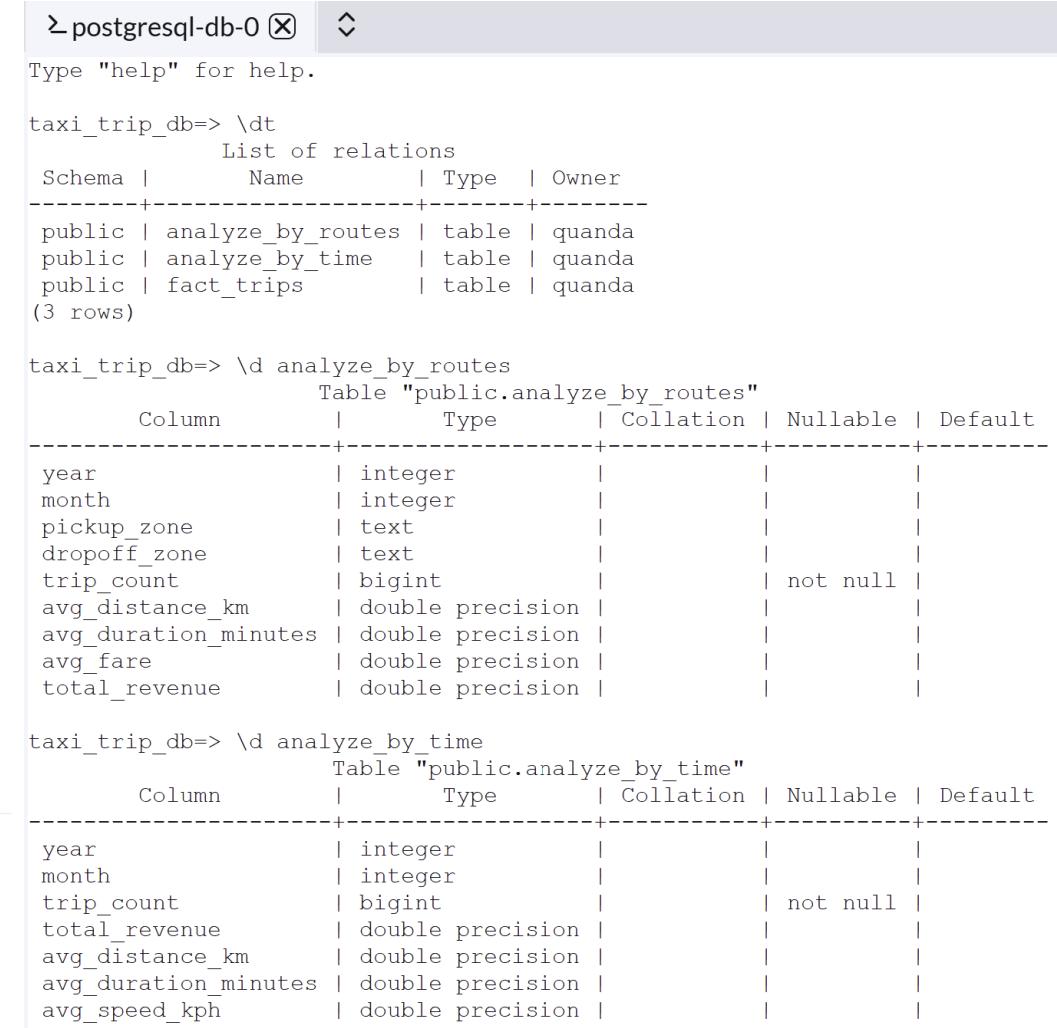
Dữ liệu thô

Dữ liệu hỗ trợ  
phân tích

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	airflow	supergroup	0 B	Jun 10 10:40	0	0 B	checkpoints
drwxr-xr-x	airflow	supergroup	0 B	Jun 10 10:06	0	0 B	models
drwxr-xr-x	airflow	supergroup	0 B	Jun 10 10:03	0	0 B	processed_data
drwxr-xr-x	airflow	supergroup	0 B	Jun 07 18:49	0	0 B	raw_data
drwxr-xr-x	airflow	supergroup	0 B	Jun 09 21:22	0	0 B	resources

Showing 1 to 5 of 5 entries

Hadoop, 2022.



postgresql-db-0

Type "help" for help.

taxi\_trip\_db=> \dt

List of relations

Schema	Name	Type	Owner
public	analyze_by_routes	table	quanda
public	analyze_by_time	table	quanda
public	fact_trips	table	quanda

(3 rows)

taxi\_trip\_db=> \d analyze\_by\_routes

Table "public.analyze\_by\_routes"

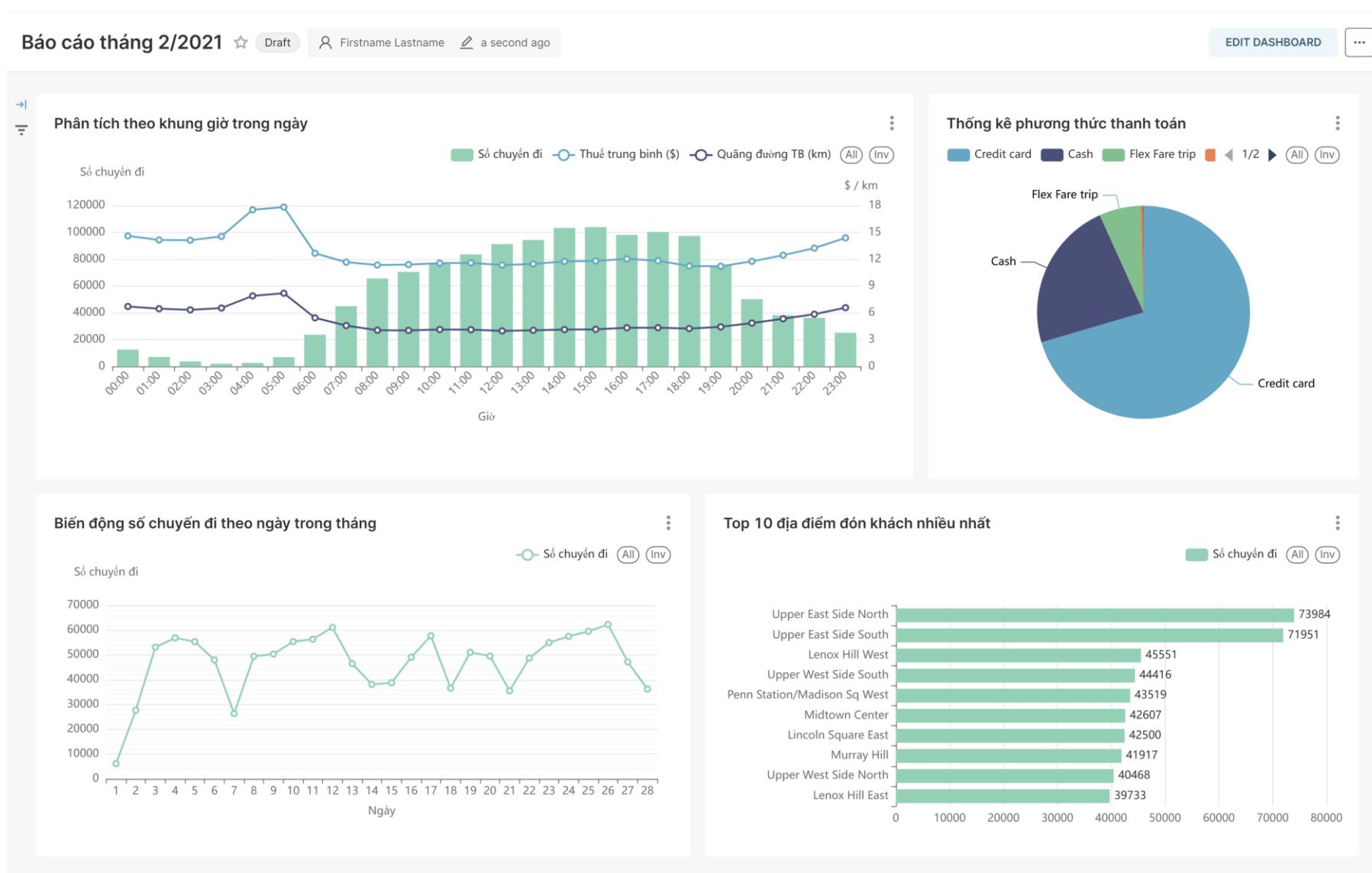
Column	Type	Collation	Nullable	Default
year	integer			
month	integer			
pickup_zone	text			
dropoff_zone	text			
trip_count	bigint		not null	
avg_distance_km	double precision			
avg_duration_minutes	double precision			
avg_fare	double precision			
total_revenue	double precision			

taxi\_trip\_db=> \d analyze\_by\_time

Table "public.analyze\_by\_time"

Column	Type	Collation	Nullable	Default
year	integer			
month	integer			
trip_count	bigint		not null	
total_revenue	double precision			
avg_distance_km	double precision			
avg_duration_minutes	double precision			
avg_speed_kph	double precision			

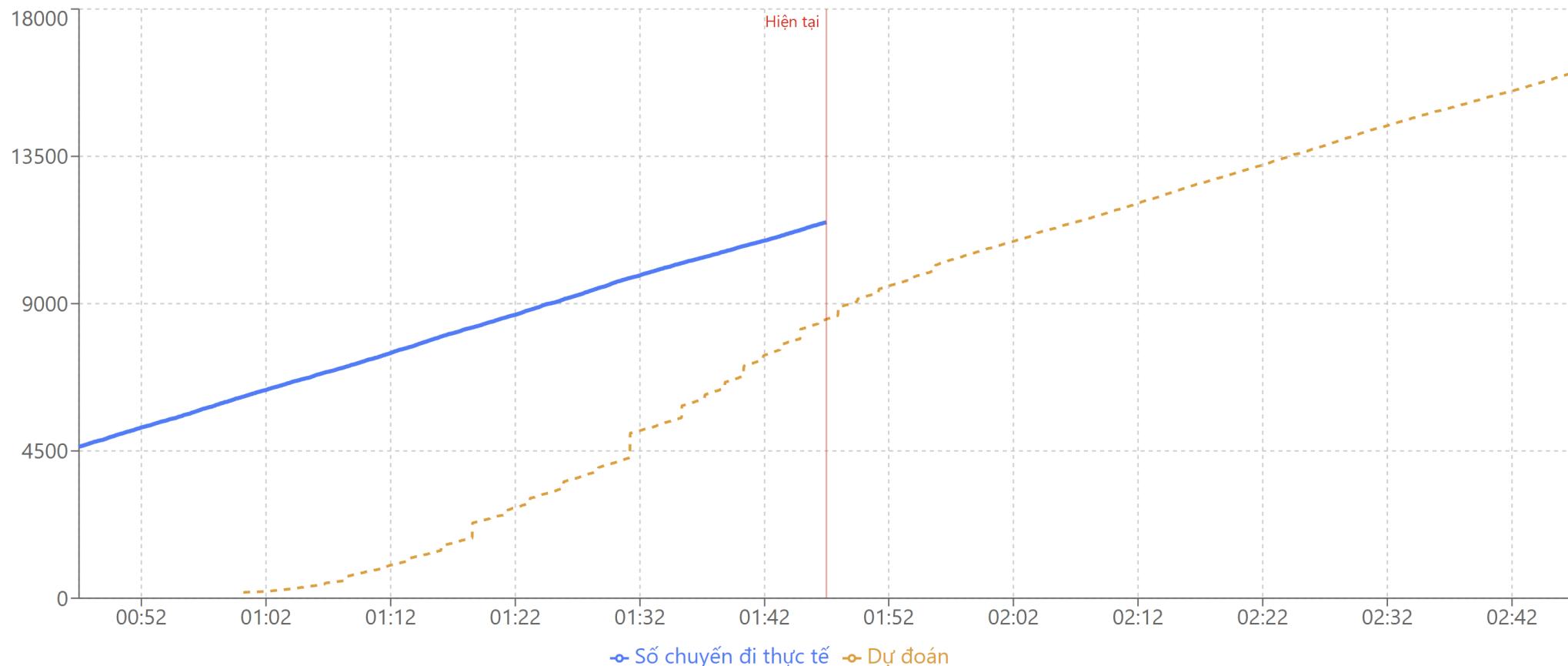
# 4. Kết quả thực nghiệm (Trực quan hóa dữ liệu)



## 4. Kết quả thực nghiệm (Trực quan hóa dữ liệu)

### Dữ liệu thời gian thực

Biểu đồ số chuyến đi theo thời gian thực ngày (01/01/2024)



# 4. Kết quả thực nghiệm (Trực quan hóa dữ liệu)

## Thống kê tuyến đường phổ biến theo địa điểm

Chọn tháng  Chọn địa điểm

Điểm đón khách	Điểm trả khách	Số chuyến	Doanh thu (\$)	Khoảng cách TB (km)	Thời gian TB (phút)
Upper East Side North	Upper East Side South	9905	114897.64	1.71	6.69
Upper East Side North	East Harlem South	4132	43933.24	1.55	5.69
Upper East Side North	Lenox Hill West	3956	48363.80	1.87	7.79
Upper East Side North	Upper West Side North	3222	41398.81	2.33	7.83
Upper East Side South	Upper East Side North	11750	133703.72	1.72	6.12
Lenox Hill West	Upper East Side North	4834	55027.01	1.81	6.37
Yorkville West	Upper East Side North	4001	39824.47	1.11	4.75
Lenox Hill East	Upper East Side North	3694	54215.96	2.11	8.29

## Thống kê dữ liệu theo tháng



# 4. Kết quả thực nghiệm (Khả năng chịu lỗi)

## Brokers

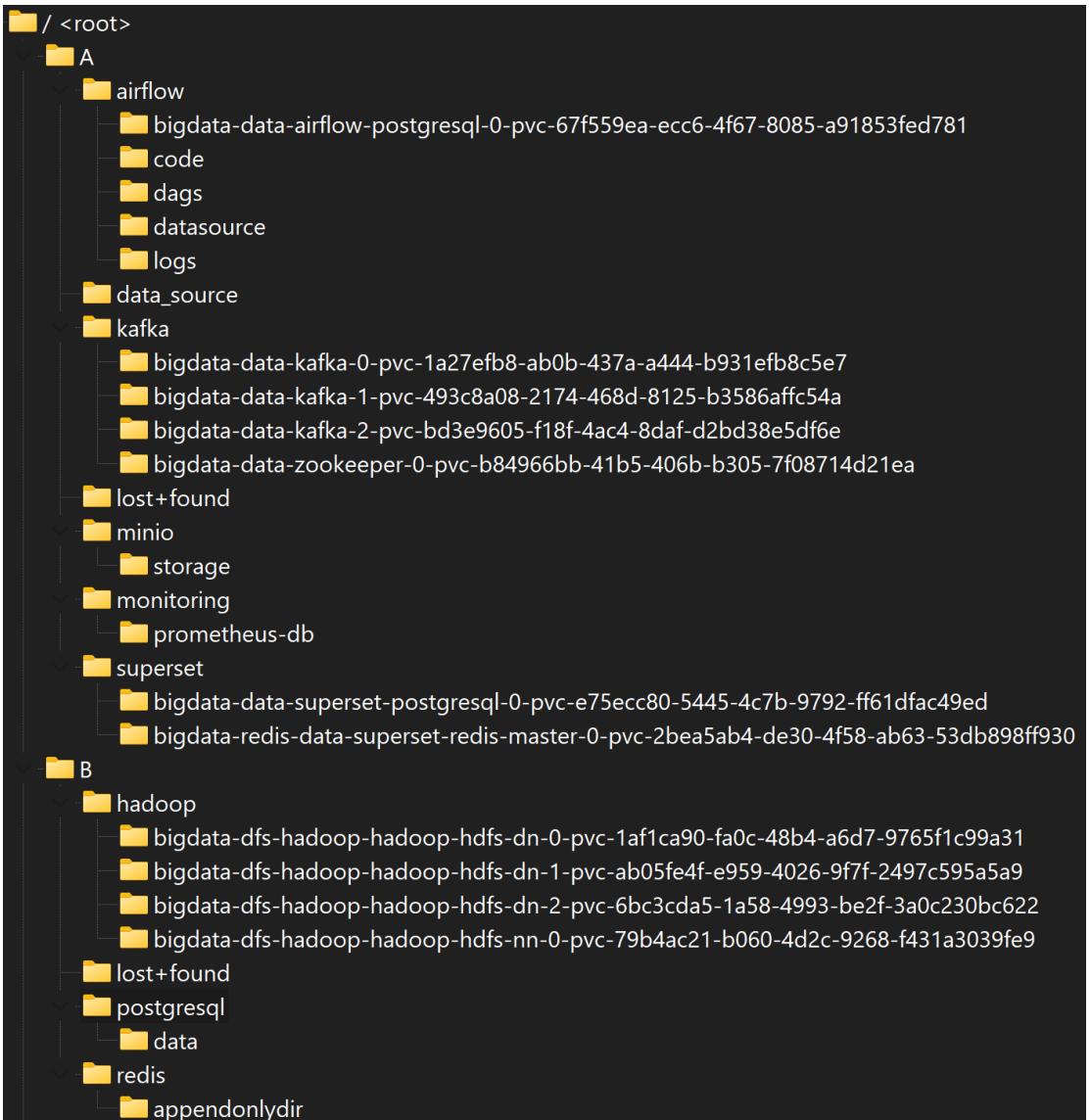
Uptime			Partitions				
Broker Count	Active Controller	Version	Online 3 of 3	URP 0	In Sync Replicas 9 of 9	Out Of Sync Replicas 0	
3	2	3.6-IV2					
0	2 MB, 3 segment(s)	-	1	-	3	9092	kafka-0.kafka.bigdata.svc.cluster.local
1	2 MB, 3 segment(s)	-	1	-	3	9092	kafka-1.kafka.bigdata.svc.cluster.local
2 <span style="color: green;">✓</span>	2 MB, 3 segment(s)	-	1	-	3	9092	kafka-2.kafka.bigdata.svc.cluster.local

## Brokers

Uptime			Partitions				
Broker Count	Active Controller	Version	Online 3 of 3	URP 3	In Sync Replicas 6 of 9	Out Of Sync Replicas 3	
2	0	3.6-IV2					
0 <span style="color: green;">✓</span>	2 MB, 3 segment(s)	-	2	-	3	9092	kafka-0.kafka.bigdata.svc.cluster.local
1	2 MB, 3 segment(s)	-	1	-	3	9092	kafka-1.kafka.bigdata.svc.cluster.local

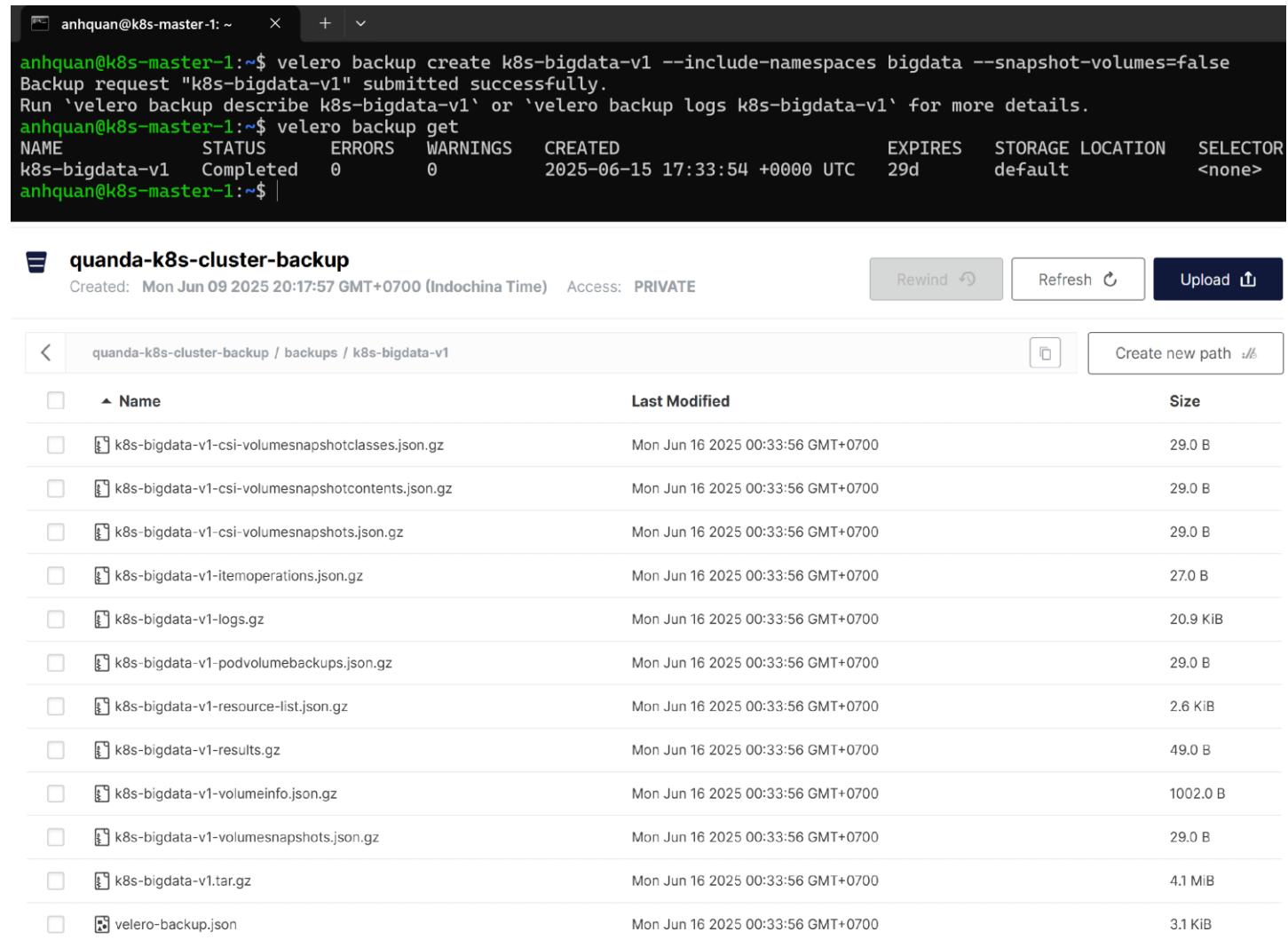
## 4. Kết quả thực nghiệm (Khả năng chịu lỗi)

- Lưu dữ liệu ra bên ngoài (NFS-Server)



## 4. Kết quả thực nghiệm (Khả năng chịu lỗi)

- Lưu dữ liệu ra bên ngoài (NFS-Server)
- Sao lưu trạng thái của cụm Kubernetes



anhquan@k8s-master-1:~\$ velero backup create k8s-bigdata-v1 --include-namespaces bigdata --snapshot-volumes=false  
Backup request "k8s-bigdata-v1" submitted successfully.  
Run 'velero backup describe k8s-bigdata-v1' or 'velero backup logs k8s-bigdata-v1' for more details.  
anhquan@k8s-master-1:~\$ velero backup get  
NAME STATUS ERRORS WARNINGS CREATED EXPIRES STORAGE LOCATION SELECTOR  
k8s-bigdata-v1 Completed 0 0 2025-06-15 17:33:54 +0000 UTC 29d default <none>  
anhquan@k8s-master-1:~\$ |

**quanda-k8s-cluster-backup**  
Created: Mon Jun 09 2025 20:17:57 GMT+0700 (Indochina Time) Access: PRIVATE  
Rewind Refresh Upload

	Name	Last Modified	Size
	k8s-bigdata-v1-csi-volumesnapshotclasses.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	29.0 B
	k8s-bigdata-v1-csi-volumesnapshotcontents.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	29.0 B
	k8s-bigdata-v1-csi-volumesnapshots.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	29.0 B
	k8s-bigdata-v1-itemoperations.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	27.0 B
	k8s-bigdata-v1-logs.gz	Mon Jun 16 2025 00:33:56 GMT+0700	20.9 KiB
	k8s-bigdata-v1-podvolumebackups.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	29.0 B
	k8s-bigdata-v1-resource-list.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	2.6 KiB
	k8s-bigdata-v1-results.gz	Mon Jun 16 2025 00:33:56 GMT+0700	49.0 B
	k8s-bigdata-v1-volumeinfo.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	1002.0 B
	k8s-bigdata-v1-volumesnapshots.json.gz	Mon Jun 16 2025 00:33:56 GMT+0700	29.0 B
	k8s-bigdata-v1.tar.gz	Mon Jun 16 2025 00:33:56 GMT+0700	4.1 MiB
	velero-backup.json	Mon Jun 16 2025 00:33:56 GMT+0700	3.1 KiB

## 4. Kết quả thực nghiệm (Khả năng mở rộng)

- Mở rộng cụm máy chủ Kubernetes
- Cơ chế nhân bản các thành phần của Kubernetes
- Các ứng dụng triển khai phân tán

## Nội dung

1. Giới thiệu đề tài
2. Thiết kế hệ thống
3. Triển khai hệ thống
4. Kết quả thực nghiệm
5. Kết luận và hướng phát triển

## 5. Kết luận và hướng phát triển

- Kết luận:

- Xây dựng thành công một hệ thống xử lý dữ liệu lớn theo kiến trúc **Lambda**, đầy đủ các thành phần từ **thu thập, xử lý, lưu trữ, trực quan hóa** đến **quản lý, điều phối**.
- Hệ thống triển khai trên môi trường **Kubernetes On-Premise**, giúp kiểm soát tối đa hạ tầng, đảm bảo khả năng mở rộng, quản lý, phục hồi hệ thống.
- Thủ nghiệm trên dữ liệu thực tế, cho thấy khả năng mở rộng tốt, hoạt động ổn định trong môi trường phân tán.

## 5. Kết luận và hướng phát triển

- **Hướng phát triển:**

- Tối ưu hóa phân chia tài nguyên, đánh giá hiệu năng xử lý, tăng cường các giải pháp bảo mật, phân quyền, hệ thống cảnh báo...
- Mở rộng nguồn thu thập dữ liệu, tích hợp thêm các mô hình học máy nâng cao, phân tích chuyên sâu
- Kết hợp với các dịch vụ Cloud, mở rộng linh hoạt với quy mô lớn



**HUST**

**THANK YOU !**